Optimizing template preparation protocol for targeted massively parallel sequencing (MPS) of the *PKD1* and *PKD2* genes: illustrative data

Ratchadaporn Chanayat[1], Ekkapong Roothumnong[2], Duangkamon Bunditworapom[2], Chanin Limwongse[2,3,4], Kriengsak Vareesangthip[2], Manop Pithukpakorn[2,4] and Wanna Thongnoppakhun[3,4]*

[1]*Graduate Program in lmmunology, Department of Immunology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*
[2]*Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*
[3]*Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*
[4]*Siriraj Center of Research Excellence in Precision Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand*

**Corresponding author: wanna.tho@mahidol.edu*

ABSTRACT

Autosomal dominant polycystic kidney disease (ADPKD) is the most common inherited nephropathy mainly resulted from mutations in *PKD1* and *PKD2* genes. Mutation analysis of *PKD1* gene is challenging because of the huge size and complicated structure which has six pseudogenes with a 97.7% identical with its exons 1-33. A traditional molecular analysis of the two genes via long-range PCR (LR-PCR) amplification followed by Sanger sequencing is still laborious and expensive. The introduction of massively parallel sequencing (MPS) allows rapid genotyping. As a prerequisite for successful MPS, sample preparation should have high-quality template libraries and sufficient yields. This study aimed to optimize the template preparation protocol for successful targeted MPS of *PKD1* and *PKD2* genes, in order to increase efficiency of ADPKD molecular diagnosis. The enriched templates, 24 LR-PCR amplicons from both genes (between 2,278 and 8,040 bp) obtained from patient samples carrying either *PKD1* or *PKD2* mutations, were applied for library preparation using QIAseq FX DNA library kit for which some particular optimizations were needed. Enzymatic fragmentation times were adjusted to be longer than that recommended by the manufacturer. In addition, amplification-free library preparation was chosen to minimize the chance of allele-dropout phenomenon usually occurring in PCR reaction especially in case of high-GC amplicons, thus the library amplification was omitted before being subjected to sequencing by the Miseq Illumina platform. Two representative data for both genes showed corresponding mutations to those identified by the previous Sanger sequencing with high coverage read depths that reflected high-quality libraries and sufficient yields. These preliminary results demonstrated that the optimized library preparation protocol for targeted MPS of *PKD1* and *PKD2* based on the templates from high-GC LR-PCR amplicons is able to detect the specified variants identified by Sanger sequencing with 100% concordance, being an effective method for simultaneous mutation analysis of both ADPKD-causing genes.

Keywords: ADPKD; *PKD1*; *PKD2*; massively parallel sequencing (MPS); long-range PCR (LR-PCR)

INTRODUCTION

Autosomal dominant polycystic kidney disease (ADPKD) is the most common inherited nephropathy characterized by progressive renal cysts in both kidneys. It affects approximately one in every 400 to 1,000 people worldwide, being the leading cause of end-stage renal disease (ESRD) in up to 10% of all cases, thus causing a major burden on public health (Spithoven *et al*., 2014). This disease is well known to be caused by mutations in two genes, *PKD1* gene discovered in 1994 account for approximately 85%, while *PKD2* identified in 1996 being responsible for the rest 15% of ADPKD patients (Barua *et al*., 2009; Harris and Rossetti 2010; Paul *et al*., 2014). Recently, genetically unresolved ADPKD-affected families were found to be caused by the *GANAB* gene which accounted for approximately 0.3% of total ADPKD cases (Porath *et al*., 2016). PKD1 patients tend to develop ESRD earlier (58

Received 11 May 2019; Accepted 6 June 2019

years) than PKD2 patients (79 years) (Boucher *et al*., 2004). Albeit as a standard tool for clinical diagnosis of ADPKD, renal imaging such as ultrasonography for renal cysts has some limitations in patients younger than 30 years-old, who yield ambiguous or inconclusive renal sonography. This would also lead to a problem in decision making of living-related kidney donors for renal transplantation of ADPKD patients. In addition, sporadic ADPKD cases (due to *de novo* mutations, mosaicism, etc.) which account for about 10–15% of patients (Reed *et al*., 2008) may be misdiagnosed based on such investigation alone. Molecular genetic testing is thus helpful for precise diagnostics of these cases. Furthermore, accurate diagnosis at a molecular level would provide the patients' opportunities to use potentially effective drugs and interventions developed for ADPKD in the present and future era of precision medicine (Lanktree and Chapman 2017).

However, molecular diagnosis of ADPKD is very challenging due to the huge size and highly complicated structure of the *PKD1* gene as well as a genetic heterogeneity with private mutations in individual families for both *PKD1* and *PKD2* genes. The *PKD1* gene located on chromosome 16p13.3 has the size around 52 kb comprising 46 exons and its main part of 5' genomic region including exons 1-33 shares 97.7% identical sequence to six pseudogenes (*PKD1P1–P6*) on the same chromosome (Harris and Rossetti, 2010). Therefore, the key step in *PKD1* sequencing is a selective amplification of the genuine *PKD1* sequence to avoid pseudogene interferences. Several designs of long-range polymerase chain reaction (LR-PCR) amplification from genomic DNA were performed with primers located on the rare mismatch sites specific to the authentic nucleotide sequences of *PKD1* (Thomas *et al.,* 1999; Thongnoppakhun *et al.,* 2000; Tan *et al.,* 2012; Tan *et al*., 2014). On the other hands, the unique regions of *PKD1* (exons 33-46) and the 68-kb *PKD2* on chromosome 4q21 which contains unique 15 exons can be analyzed directly. Sanger sequencing is still the gold standard for a final mutation analysis of both *PKD1* and *PKD2* genes regardless of whether nested PCRs are amplified from the specific LR-PCR amplicons (Rossetti *et al*., 2001; Tan *et al.,* 2012) albeit being laborious, time consuming and expensive. Alternative method with lower cost such as denaturing high performance liquid chromatography (DHPLC) may be used as an additional screening (Rossetti *et al.,* 2002), but this approach has to be performed in batch mode for cost and time saving.

A recent high-throughput technology, massively parallel sequencing (MPS) commonly known as next-generation sequencing (NGS) is capable of simultaneous sequencing of tremendous numbers of DNA strands up to a whole genomic sequence by generating millions reads of sequences. The implementation of MPS in molecular diagnostics enables a more cost-effective method than previous ones with a higher diagnostic yield (Renkema *et al*., 2014). It was suggested that MPS of the *PKD1* and *PKD2* genes could substitute for Sanger sequencing as the standard approach for genetic testing of ADPKD with high sensitivity and significant reduction of turnaround time and costs compared with Sanger sequencing (Tan *et al*., 2014; Trujillano *et al*., 2014; Ranjzad *et al*., 2018). However, the application of MPS to routine diagnosis laboratories would require validation of the method.

Targeted MPS of the *PKD1* and *PKD2* genomic regions can be performed through the initial target enrichment of both genes via either LR-PCR or selective captures with specific probes. To reduce the costs of sequencing, a varying number of DNA samples can be pooled for genotyping in one experiment by using unique barcoding of each sample for enabling the tracing of individual sample data (Renkema *et al*., 2014). The sample identification allows sample multiplexing from a batch of patient samples that could be analyzed within a single MPS run (Moorthie *et al*., 2011).

To achieve successful MPS sequencing, library preparation is a critical step as a low-quality library often yields very poor data. The library preparation process involves multiple handling steps, including template fragmentation, end-repair, adapter ligation and PCR amplification (van Dijk *et al*., 2014a). This process is generally time consuming, being able to result in sample loss and handling errors. Some newly commercially available library preparation kits may improve the process to save times and reduce errors. Quality control (QC) are also important to ensure the integrity of template at each step of the process. However, the sample preparation still remains a technical challenge, requiring some degree of improvements for the particular template preparation protocols to enhance the sequencing data quality (van Dijk *et al*., 2014b).

Here, we present an alternative method of template preparation from the LR-PCR amplicons of *PKD1* and *PKD2* genes with optimized library preparation by increased fragmentation time, more frequent QC steps and PCR omission. The

representative data obtained show the complete concordance with the previous Sanger sequencing, reflecting the high quality MPS data appropriate for clinical diagnostics of ADPKD.

MATERIALS AND METHODS

Studied samples and ethics statement

The samples used in this study were ADPKD patients' DNA samples stored for molecular genetic study at our Molecular Genetics Laboratory, Division of Medical Genetics Research and Laboratory, Research Department, Faculty of Medicine Siriraj Hospital, Mahidol University. This study was approved by Siriraj Institutional Review Board (Si-IRB), COA number: Si 439/2008. All subjects gave written informed consents for blood collection. The samples selected for study comprised the first group of 10 known ADPKD cases with known various types of mutations in both *PKD1* and *PKD2* genes as characterized by traditional methods, LR-PCR or standard PCR followed by direct Sanger sequencing as well as multiplex ligation-dependent probe amplification (MLPA). In addition, the second group of 9 unknown ADPKD cases whose causative mutations have not yet been identified and one normal Thai volunteer who gave written informed consent were also recruited in this study.

Genomic DNA quantity and quality assessment

The genomic DNA quantity and quality were assessed by UV spectrophotometer, Nanodrop® ND-1000 (Thermo Fisher Scientific, IL, USA) and 0.8% agarose gel electrophoresis, respectively.

Long-range PCR (LR-PCR) of the entire *PKD1* and *PKD2* genes

The set of nine primers (Liu *et al.*, 2014) and a newly designed primer pair (more extended 3' untranslated region, UTR) were used to amplify 10 LR-PCR amplicons throughout the *PKD1* gene, with following sizes (bp) from 5' to 3' respective regions as 2,278, 4,041, 3,893, 4,391, 4,350, 3,301, 3,916, 2,632, 2,909 and 4,170. Meanwhile, the *PKD2* gene was amplified with all newly designed primers into 14 LR-PCR amplicons overlapping each other with at least 500 bp throughout the entire gene, with following sizes (bp) from 5' to 3' regions as 8,040, 6,425, 6,017, 6,095, 6,235, 6,216, 6,201, 6,191, 6,098, 6,122, 6,398, 6,440, 5,990 and 7,529, respectively. Primer sequences for LR-PCR amplification of *PKD1 and PKD2* genes which were newly designed in this study are shown in Table 1. LR-PCR products from the *PKD1* gene have high GC contents ranging from 60.1 to 70.2% (Liu *et al.*, 2014). We used PrimeSTAR GXL DNA Polymerase (Takara Bio Inc., Japan) for all LR-PCR amplifications, except for the 2,278-bp and 2,909-bp fragments for which Phusion Flash High-Fidelity PCR Master Mix (ThermoFisher Scientific, USA) and PrimeSTAR Max DNA Polymerase premix (Takara Bio Inc., Japan) were applied instead. The obtained amplicons were stored at 4°C for the library preparation step.

Library preparation according to the manufacturer's protocol

All the 24 LR-PCR amplicons were subjected to library preparation with QIAGEN® QIAseq FX DNA Library Kit (QIAGEN Inc., USA) following the manufacture's protocol. Briefly, these amplicons were purified by Agencourt AMPure® XP beads (Beckman Coulter, USA) according to the product manual, then quantified by Quant-iT™ PicoGreen® dsDNA assay (Quanti-iT; Invitrogen, USA) measured with Infinite®200 PRO microplate reader (TECAN, Austria) or Qubit dsDNA HS (high sensitivity, 0.2 to 100 ng) Assay Kit detected with Qubit™ 3.0 Fluorometer (Life Technologies, USA) following the manufacturers' instruction. For each sample, individual amplicons were diluted to 10 ng/µL and pooled together, 100 ng of individual pooled amplicons from each patient's sample were subjected to enzymatic fragmentation to generate amplicon sizes of approximately 250 bp and directly end-repaired within a single tube. After size selection, the obtained libraries were amplified and assessed for the quality by Agilent 2100 Bioanalyzer (Agilent Technologies Inc., USA) using the High Sensitivity DNA kit (Agilent Technologies Inc., USA) following the manufacturer's instruction just only once before moving to the massively parallel sequencing step.

Optimization of library preparation protocol

The following modifications were performed in this study:

Increasing the reaction time of enzymatic fragmentation

As recommended by the manufacture's protocol, 100 ng of input pooled amplicons should be incubated with FX Enzyme Mix at 32°C for 16 min. However, according to the high-GC content of the LR-PCR amplicons especially from the *PKD1* gene, we decided to optimize the reaction time of enzymatic fragmentation by increasing of 2 and 4 min (to be 18 and 20 min, respectively) to achieve the expected size range, in parallel with the use of 16 min following the

manufacturer's protocol. At this step, the first two amplicon-pooled samples were randomly selected to obtain preliminary results that would be optimal and applied for all the samples studied. After this step, these fragmented amplicons were ligated with adapters and barcodes.

Table 1. Primer sequences for LR-PCR amplification of *PKD1 and PKD2* genes (showing only newly designed primers in this study)

No.	Covered Region	Primer Name	Nucleotide sequence (5'-3')	Primer size (bp)	LR-PCR product size (bp)	%GC of PCR products
For *PKD1* gene						
1.10	Ex 40-TSC2	PKD1-10F	TGTTTCAGCACTCTACCCAGACCCT	25	4,170	67.5
		PKD1-10R	GTGACCACCAAGTCTCCCCAGACAT	25		
For *PKD2* gene						
2.1	c.1-2316 – Exon 1	PKD2-1F	ACTAGGCCTATTCTCCAGAGGACCC	25	8,040	45.3
		PKD2-1R	CCCTGAGCTCCTTGAAGGTCGAGA	24		
2.2	Intron 1	PKD2-2F	TTCTCCCACCCAAATTGCCAGGAGT	25	6,425	42.9
		PKD2-2R	CAACCTGGTCCTTGCATGTTGT	22		
2.3	Exon 2	PKD2-3F	GTATAAATATTAAGAATCATGGCTG	25	6,017	38.9
		PKD2-3R	CTTTGGCAAGAATAACTTCTTACTA	25		
2.4	Intron 2a	PKD2-4F	GGCTCTTTGGTTGGAGTTCACTT	23	6,095	42.9
		PKD2-4R	CCTCCAGTCATAGTTGAAGGCTT	23		
2.5	Intron 2b	PKD2-5F	GGACAGGAAGAGCCCTTTTAAGA	23	6,235	39.1
		PKD2-5R	AGGAGAAGGGTAGATAATAATGATC	25		
2.6	Exon 3-4	PKD2-6F	GGTGCTACATAACCAAGGGAATCTG	25	6,216	41.6
		PKD2-6R	GTTCCTTTGTTCTAAAGCTCTGGAT	25		
2.7	Exon 5	PKD2-7F	CGAAATGGAACCGCGTAAGTGTCT	24	6,201	42.3
		PKD2-7R	CCTTCAGGCTATGTGCATTAAG	22		
2.8	Exon 6	PKD2-8F	CTGGTCCAAGTTCATCTCAATGCTG	25	6,191	37.9
		PKD2-8R	GACTATTGCTCAGGACAAATGGC	23		
2.9	Exon 7	PKD2-9F	ACATTTTCTTCCCATAGCCCAGTGA	25	6,098	42.1
		PKD2-9R	TGCCGTGAACTGTGACTAGTGATGC	25		
2.10	Exon 8-9	PKD2-10F	GGTGCTGCCTGAGAATTACTGAAGT	25	6,122	38.6
		PKD2-10R	GAAACAGATCAAAGTCTGGGACCTC	25		
2.11	Exon 10-11	PKD2-11F	TGATCTCAGGCCACCTAGTTTTC	23	6,398	38.2
		PKD2-11R	GCTGATGTTCATGTTCGGTCAGTTC	25		
2.12	Exon 11-13	PKD2-12F	AGGTCAGGTCTAAAGTTCGCTAATG	25	6,440	37.9
		PKD2-12R	TACAGAATAAAGAGACAGAGATTGC	25		
2.13	Intron 13-Exon 14	PKD2-13F	TACTGTGGTCCCCTTGTACATGATG	25	5,990	42.6
		PKD2-13R	GTTTACCCCAGGAAGCTAGTTGA	23		
2.14	Exon 14 - *4083	PKD2-14F	TTAACAGGATTAGAACCTCTGCAAC	25	7,529	40.0
		PKD2-14R	GTAAAAGAACCAGCACAGTTCGTAA	25		

Library preparation without PCR amplification step

The PCR-based library amplification step was omitted because of the sufficient (100 ng) amount of the input DNA amplicons and, importantly, to reduce the amplification errors introduced especially in the GC-rich templates.

More frequent assessment of library quality

We assessed the quality of the prepared libraries using Agilent 2100 Bioanalyzer more frequently than that recommended by the manufacturer (only once before MPS step), by evaluating in many steps including after fragmentation, after adapter ligation (before library amplification) and after library amplification (if any).

Sample pooling and massively parallel sequencing (MPS)

Equivalent amounts (4 nM) of individual sample libraries, as normalized libraries from a total of 20 samples including the 10 known ADPKD cases, 9 unknown ADPKD cases and one normal control subject were pooled together before going to the next step. Seven picomolar of multiplexed libraries were

loaded on the Illumina Miseq sequencer (Illumina Inc., USA) using Miseq Reagent Kit v2 (Illumina Inc., USA). The cluster generation and paired-end sequencing with 250 bp read length 500 cycles were performed automatically within the Illumina Miseq sequencer following the manufacturer's instruction.

Bioinformatics data analysis and variant validation

The FASTQ files were then analyzed by SeqNext module of Sequence Pilot software, version 4.4.0 (JSI Medical systems, Germany) using NM_001009944.2 for *PKD1* and NM_00297.3 for *PKD2* as the reference sequence to generate the Region of Interest (ROIs) for sequence alignment and variant detection. The workflow includes mapping, alignment, variant calling, QC-filtering, visualization and filtering. The generated sequencing reads were mapped to defined ROIs. Briefly, the candidate variants which account for 20% of both forward and reverse reads were marked as "distinct variants". The high frequency variants were discarded as filter-out variants with allele frequency ≥1% according to various databases including 1000 Genomes (http://www.internationalgenome.org/) and Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org) as well as Genome Aggregation Database, gnomAD (http://gnomad.broadinstitute.org). The variant annotation and classification were assessed with a combination of allele frequencies and multiple *in silico* prediction tools, for example, Polyphen-2 (http://genetics.bwh.harvard.edu/pph2), SIFT (http://sift.bii.a-star.edu.sg/), and MutationTaster (http://www.mutationtaster.org/). The obtained variants were compared with the previously identified mutations by Sanger sequencing of each case.

RESULTS

LR-PCR amplicons of *PKD1* and *PKD2* genes

According to the large size of both genes and the presence of six pseudogenes for the *PKD1* gene, LR-PCR was applied to specifically amplify only the authentic *PKD1* entire coding sequence and the whole *PKD2* coding region. The *PKD1* gene was generated into 10 amplicons as shown in Figure 1 (Top). For the *PKD2* gene amplification, 14 LR-PCR amplicons produced by newly designed primers in this study are shown in Figure 1 (Bottom). These LR-PCR amplicons were quantified by PicoGreen and prepared to 10 ng/μL. Ten microliters of each amplicons (10 ng/μL) were pooled together followed by AMPure XP beads purification at a ratio of 1.8x and these pooled amplicons were then subjected to MPS library preparation.

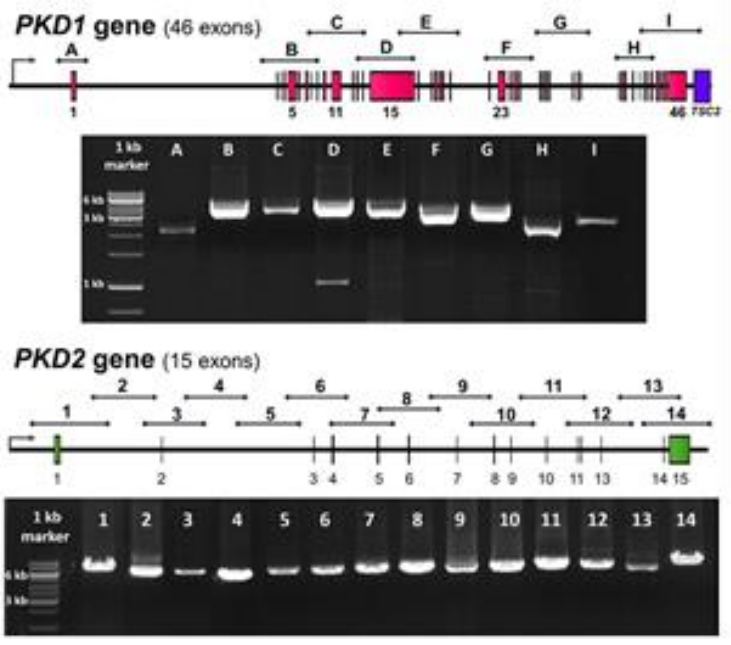

Figure 1. Representative results of LR-PCR amplifications for *PKD1* (Top) and *PKD2* (Bottom) genes. Pink, blue and green boxes are exonic regions of *PKD1*, the adjacent *TSC2* and *PKD2* genes, respectively. For each studied sample, LR-PCR amplicons were amplified into 10 amplicons for *PKD1* (2.2-4.5 kb) and 14 amplicons for *PKD2* (6.0-8.0 kb).

Optimization of MPS library preparation from the LR-PCR amplicons

The QIAGEN® QIAseq FX DNA Library Kit was chosen for NGS library preparation in this study. No previous information about the use of this kit for LR-PCR amplicons was available. An optimized condition for this study was thus required for this relatively new commercially available kit. Two amplicon-pooled samples were initially tried before application of the optimized library preparation protocol to all 20 studied samples.

The expected size after amplicon fragmentation is 250 bp and these amplicons were then ligated to the 120-bp adapter. Therefore, the achieved library should show a median of distribution size around 370 bp (the size of the fragmented amplicons plus 120 bp). However, the initially prepared libraries revealed unexpected sizes of larger libraries (Figure 2, Top). The fragmentation time provided by the manufacturer (16 min) might be not enough for LR-PCR amplicons in this study. Increment of the reaction time by 2 and 4 min (to be 18 and 20 min) was thus tried followed by assessing the size distribution of libraries for each fragmentation time. This was an additional assessment of the library quality, being more frequent than that in the manufacturer's recommendation which suggested only once just before loading to the MPS instrument.

The results of both reaction time increments gave an expected size of amplicons around 370 bp, the large size amplicons are absolutely removed and the results of two durations (18 and 20 min) were quite not different from each other (Figure 2, Bottom). The 4-min increment (20 min) seemed to give higher yield of expected fragment sizes as indicated by higher fluorescent unit. After adapter ligation, the size distributions were assessed again, it was found that the expected size of libraries was still obtained although the average library size was slightly larger compared with the PCR-amplified libraries (Figure 2, Bottom). According to these results, LR-PCR amplicon libraries were then prepared by increasing the fragmentation reaction time to 20 min followed by barcode and adapter ligation. The library amplification step was performed and the amplified libraries were also checked (Figure 2, Bottom). However, skipping of the library amplification step was another modification applied in this study to avoid the allele-dropout phenomenon which usually occurs during PCR process, the PCR-free adapter-ligated libraries were thus directly subjected to the MPS step.

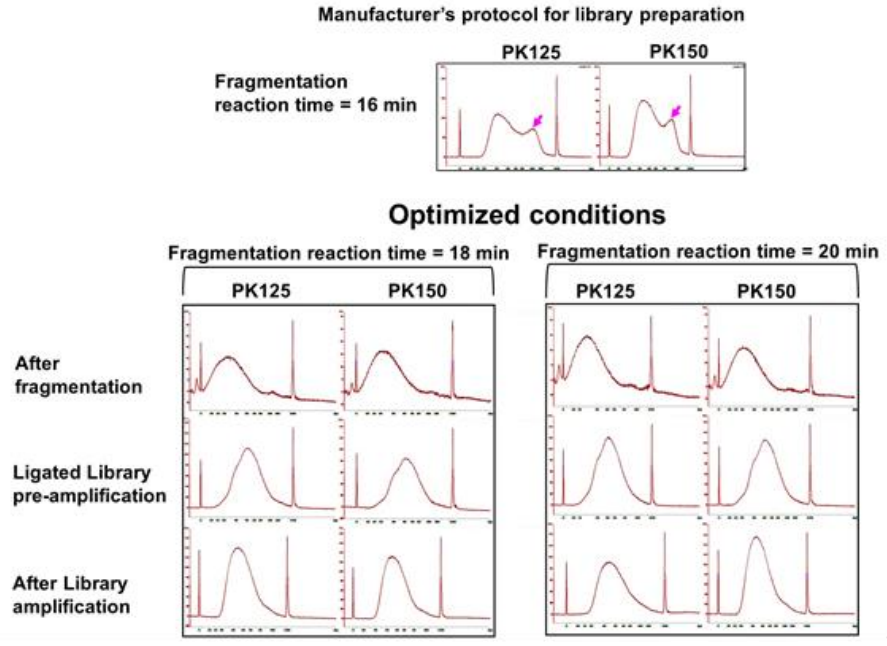

Figure 2. Optimization of fragmentation reaction time for LR-PCR amplicons using QIAGEN® QIAseq FX DNA Library Kit represented by 2 samples: PK125 and PK150. Obtained libraries according to the manufacturer's protocol and quality check by fragment analysis (Agilent 2100 Bioanalyzer), indicating incomplete fragmentation due to the existence of libraries with larger size than expected (pink arrow) (Top). Elimination of large library populations by increment of reaction times to 18 and 20 min (Bottom).

Data analysis of the targeted massively parallel sequencing (MPS) in the pooled samples

The pooled libraries of 20 samples comprising 10 known ADPKD cases, 9 unknown ADPKD cases and one normal control subject were analyzed through the LR-PCR-based targeted MPS for *PKD1* and *PKD2* genes in a single flow cell of Miseq sequencer.

In a single run, 3 Gb of sequencing data was achieved with the cluster density of 438K/mm^2. Up to 100% of bases were mapped back to the ROI with more than 30x. From the single batch of 20 samples, two known ADPKD cases were selected to present in this paper as preliminary data for each gene, PK125 (carrying *PKD1* mutation) and PK150 (having *PKD2* mutation). By this approach, both samples achieved high read depth across all regions of *PKD1* and *PKD2* genes as shown in Figure 3. Up to 98% of all sequence reads could be mapped to the reference genome. A hundred percent of targeted regions of *PKD1* and up to 97% of *PKD2* were covered with more than 30 reads. Focusing on all exons of *PKD1* gene had minimum read depth of 422x and maximum read depth of 1,332x (Figure 3, Top) and *PKD2* exonic regions were covered with 506–1,661-fold depth (Figure 3, Bottom). These highly read depths suggested that calling of distinct variants was feasible for all exons of both genes.

Variants identified by this MPS compared with Sanger sequencing variants: 2 representative cases

Two representative samples for either *PKD1* or *PKD2* genes, PK125 and PK150 were previously genotyped by Sanger sequencing. The targeted MPS based on LR-PCR of both genes which was optimized as mentioned above in this study could give the completely concordant results. PK125 was found to have a heterozygous in-frame indel in exon 40 of the *PKD1* gene, c.11348_11385delinsAGTCC (p.Val3783_Trp3795delinsGluAla), while PK150 having a heterozygous nonsense mutation in exon 13 of *PKD2* gene, c.2407C>T (p.Arg803Ter) as shown in Figure 4 and 5, respectively.

Figure 3. Read depth and coverage results of each LR-PCR amplicon of *PKD1* (Top) and *PKD2* (Bottom) genes represented by 2 samples: PK125 and PK150. The average read depth of each fragment (focusing on exonic covered region) is illustrated above each amplicon. The plots were generated by Integrative Genomics Viewer (Top) and JSI SeqPilot software (Bottom).

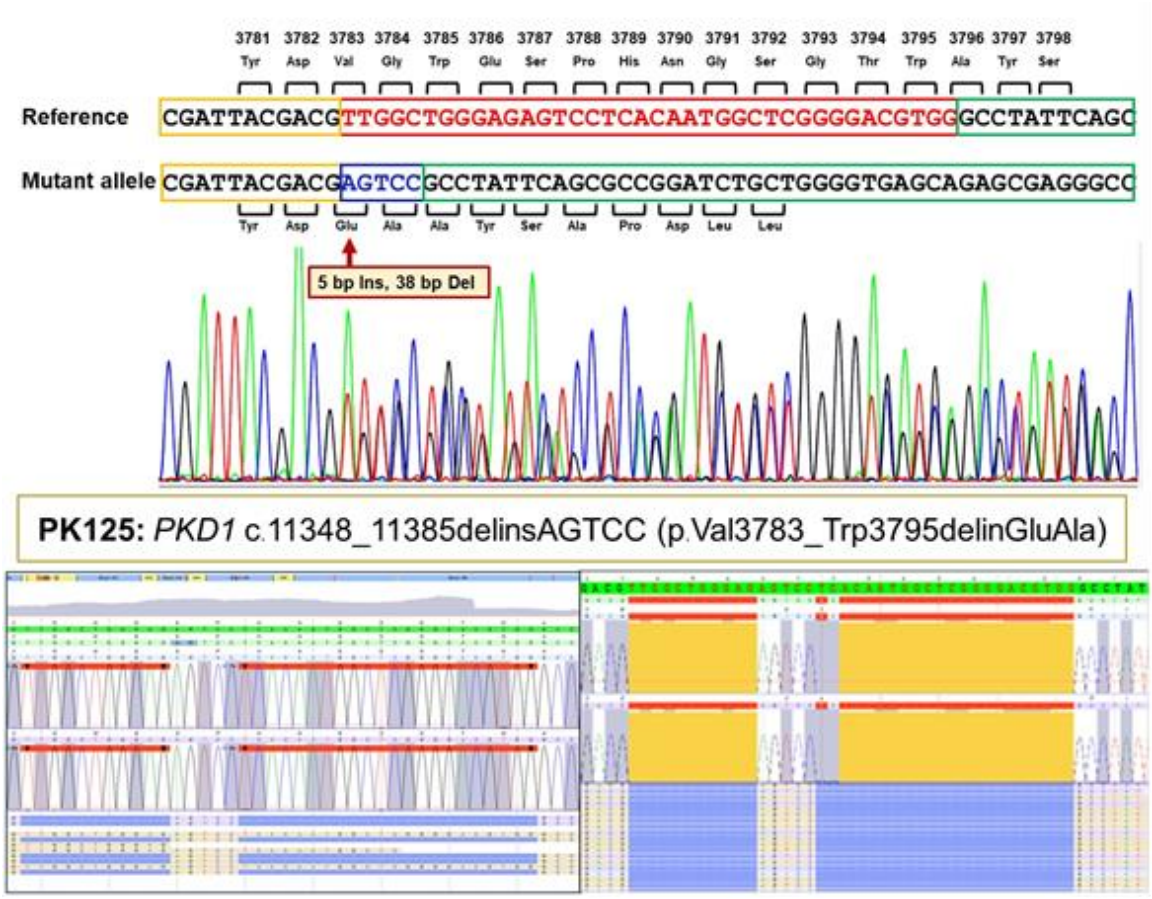

Figure 4. Illustrative MPS data for a *PKD1* mutation found in PK125 patient who has a heterozygous in-frame indel in *PKD1* exon 40, c.11348_11385delinsAGTCC (p.Val3783_Trp3795delinsGluAla) showing the concordance results. The previous Sanger sequencing result (Top) is used as reference for comparison with the mutation identified by the present targeted MPS approach (Bottom).

In addition to the pathogenic mutations in both representative cases (PK125 and PK150) which could be identified concordantly by the two methods, this targeted MPS approach also detected several additional variants in both genes. In the case of PK125, there were six *PKD1* variants (c.10168-21A>C, c.11016+87dupC, c.11156+164G>A, c.*706C>T, c.*1069_*1072dupTGAC and c.*1080A>G) and two *PKD2* variants (c.596-3755T>C and c.710-7847C>T). PK150 was found to have only one for each gene, *PKD1* c.4472G>T and *PKD2* c.*1999C>T. However, the previous Sanger sequencing found only a few of these variants because the previous primer design did not cover the deep intronic regions and the more extended 3' UTR region as the primers designed for this targeted MPS.

DISCUSSION

Targeted MPS based on LR-PCR amplicons of *PKD1* and *PKD2* genes: Design rationale

The size and structure complexity of *PKD1* and *PKD2* genes especially the existence of six pseudogenes of *PKD1* make the analysis of both genes cumbersome. The gold standard for ADPKD diagnosis used by general molecular laboratories is LR-PCR followed by direct Sanger sequencing which are still laborious, time consuming and costly. Therefore, MPS was applied instead of Sanger sequencing for analysis of LR-PCR amplicons from *PKD1* and *PKD2* genes. The use of LR-PCR amplicons as the primary target enrichment has several advantages as high sensitivity, high specificity, reproducibility and uniformity (Moorthie *et al*., 2011), being appropriate for the study of just these two genes.

MPS have recently become popular and revolutionized genetic research and genetic diagnosis in many inherited disorders (Park *et al*., 2016). This technology allows high-throughput sequencing simultaneously, resulting in shorter turnaround time while being cost-effective. Library preparation is a critical part of the MPS workflow. Successful sequencing requires the generation of high quality

libraries of sufficient yield and quality, especially the beginning step of library preparation. This involves shearing DNA input to generate expected amplicon size which could be proceed by either mechanical shearing, enzymatic fragmentation or chemical fragmentation (Bowers *et al*., 2015). There are numerous differences advantages and limitations for all 3 options. But for the last 5–10 years, mechanical fragmentation with Covaris technology which uses acoustic shearing has been the gold-standard for DNA fragmentation (Knierim *et al*., 2011). It avoids most GC bias and provides relatively consistent fragment sizes thus good quality of library was produced. However, the entire process is time- and labor-intensive, as well as expensive with a high cost for new instrumentation and single-use consumables. There is also the issue of sample loss during multiple transfer steps between tubes.

Alternative library preparation methods currently available include a chemical method and enzymatic method such as tagmentation or fragmentase which is easy for user and no need of experienced person to deal with (Head *et al*., 2014). This was reasonable for choosing an enzymatic fragmentation-based library preparation kit for this study which was QIAGEN® QIAseq FX DNA Library kit.

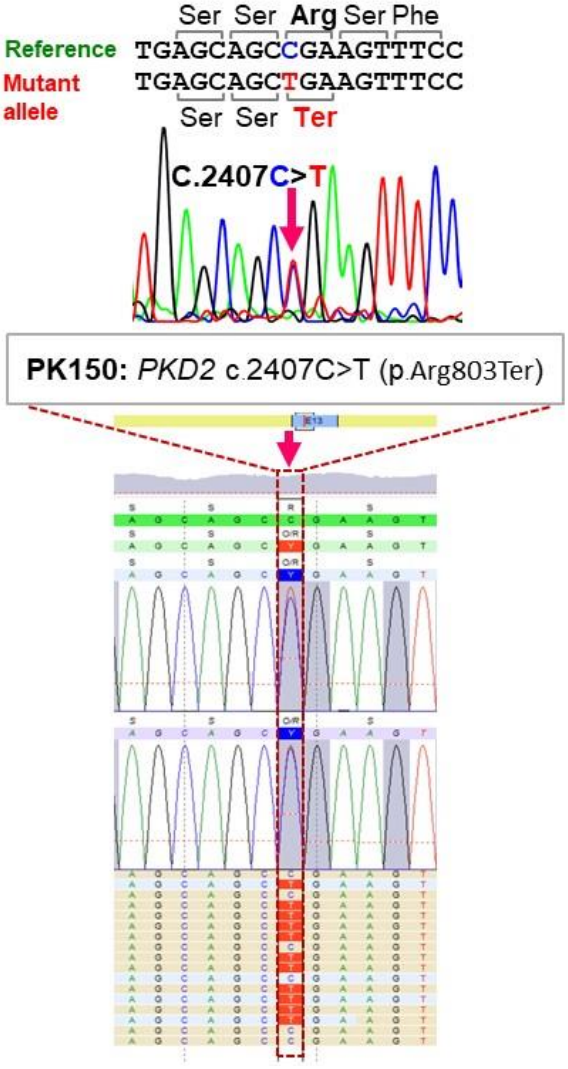

Figure 5 Illustrative MPS data for a *PKD2* mutation found in PK150 patient who has a heterozygous nonsense *PKD2* mutation in exon 13, c.2407C>T (p.Arg803Ter) showing the concordance results. The previous Sanger sequencing result (Top) is used as reference for comparison with the mutation identified by the present targeted MPS approach (Bottom).

There were many research groups utilized the MPS technology to develop efficient diagnostics for ADPKD. According to the presence of six pseudogenes which shared a 97.7% identity to exons 1-33 of *PKD1* gene, Rossetti *et al*. (2012) firstly developed NGS for high-throughput mutation screening based on specific LR-PCR amplicons amplified by primers located on the rare mismatched sites to discriminate pseudogenes. According to the low sensitivity and long turnaround time resulted from inefficient sample-pooling strategies of the former study, Tan *et al* (2014) reported a tailored NGS-based genotyping approach for ADPKD with rapid turnaround time and improved sensitivity. They individually barcoded LR-PCR amplicons of each patient before pooling instead of DNA/amplicon pooling of each sample before library construction as performed by Rossetti's group. This study also reported the ability to identify variants missed by Sanger sequencing. Technologist's error and allele dropout during PCR amplification step were described to be the causative of this case.

A capture-based approach was also developed to capture the targeted exons by commercially specific probes. The biased capture for GC-rich regions still results in insufficient coverage of variant calling (Qi *et al*., 2013). In the case of *PKD1* gene which require locus-specific amplification to exclude the interference of six pseudogenes, it is thus difficult to design capture probes, short oligonucleotides to avoid capturing pseudogenes. For these reasons, a simple ADPKD analysis approach based on targeted MPS using 24 LR-PCR amplicons (between 2,278 and 8,040 bp) of *PKD1* and *PKD2* genes was developed as a model for analysis of complicated genes and further clinical use.

Optimizing library preparation for LR-PCR (longer fragmentation time, frequent QC check, PCR-free)

Generation of MPS libraries was performed using a QIAGEN® QIAseq FX DNA Library Kit which included enzymatic fragmentation of DNA template, adaptor/barcode ligation and optional library amplification. As a quite newly available in the market, this library preparation kit has no previous information about the optimal use for LR-PCR from human genome with high-GC content and sophisticated structure as *PKD1* gene. Optimization of this commercial kit would provide useful information as a guidance for further applications. To the best of our knowledge, this is the first evaluation of this kit for its use in LR-PCR based targeted MPS for ADPKD diagnostics.

In this study, the QIAseq FX DNA Library kit was optimized for the enzymatic fragmentation reaction time. As the use of 100 ng input DNA (LR-PCR amplicons) to get the expected fragment peak size of 250 bp (plus 120-bp ligated adapter to be a median size distribution around 370 bp), it was recommended by the kit to use fragmentation time at 32°C for only 16 min. However, the optimal fragmentation was obtained by increment of reaction times to either 18 or 20 min, and the latter was preferably chosen for further study. Quality control check by fragment analysis at this step is also important. The modified fragmentation time to be longer as 20 min and monitoring of library yield and quality enabled the generation of sequencing libraries with sharper and more symmetric insert size distributions (Figure 2**)**. It was previously found that read pairs generating from longer DNA fragments tended to have worse sequencing quality because of the disproportionately often failure to pass quality filtering which was especially strong for libraries with high-GC content (Huptas *et al*., 2016). Therefore, the smaller insert sizes from the high-GC content genes (*PKD1* and *PKD2*) obtained from the 20-min fragmentation time should give better results than the use of 16- or 18-min fragmentation. As being designed to save time and prevent errors, library preparation by the QIAseq FX kit combined fragmentation and adapter ligation step within a single tube reaction.

An optional PCR amplification of library DNA is recommended in cases of low input DNA such as 6 library amplification cycles for 100 ng input DNA. However, the amounts of DNA before and after library amplification were similar (Figure 2). To minimize the chance of allele-dropout phenomenon or uneven read coverage especially across GC-rich regions in *PKD1* and *PKD2* LR-PCR amplicons, the library amplification step by PCR was therefore omitted, PCR-free libraries were then subjected to further step. When performing PCR-based diagnosis, "allelic dropout" can cause misdiagnosis due to missing of some variants or mutations. There are various mechanisms that result in this phenomenon for example, poor quality sample, unbalanced locus specific amplification by PCR reaction and high GC-content amplicons (Stevens *et al*., 2017). It was suggested that a reduction in the number of unnecessary PCR cycles or omitting of PCR amplification can significantly reduce the incidence of duplicate sequences resulting in increased data quality of MPS with improved read mapping and variant calling (Kozarewa and Turner, 2011). Therefore, our

library preparation from the LR-PCR amplicons of *PKD1* and *PKD2* genes were PCR-free.

MPS data analysis of the pooled samples of two illustrative cases

The MPS data obtained in the two illustrative cases, PK125 and PK150 showed the high read depth of 422-1,332x for *PKD1* and 506-1,661x for *PKD2* (Figure 3), indicating that the optimized library preparation protocol could generate a considerably uniformly high coverage, non-biased source of DNA across both genes even within sequences with high GC contents.

The results from the two illustrative samples for both *PKD* genes could infer the overall MPS data quality of all 20 samples in the same pool. After the enzymatic fragmentation step of DNA templates (LR-PCR amplicons) using the QIAGEN® QIAseq FX DNA Library Kit, two indices were added (dual-barcoded ligation) to each DNA fragment. According to the kit design, up to 24 or 96 uniquely indexed samples could be pooled and sequenced together in a single run on the Illumina Miseq sequencer, however, to balance the sample throughput, we decided to use 20 samples in a single pool as an optimal number of samples in terms of analytical sensitivity (sufficient sequencing depth). This is a cost-effective method by maximizing the throughput of a single run for molecular diagnosis of ADPKD based on the targeted MPS. The sample pooling carefully used equivalent amounts (4 nM) of individual sample libraries to avoid the over- or under-representation of some samples, minimizing an unequal sample bias resulting in coverage uniformity and limited sequencing biases.

In spite of confined performance of MPS in GC-rich regions of the human genome, we were able to detect the *PKD1* mutation in exon 40, c.11348_11385delinsAGTCC (p.Val3783_Trp3795delinsGluAla) (Figure 5), a region considered as challenging due to the highest GC content (70.2%) of the 2,909-bp LR-PCR product (Liu *et al*., 2014) amplified from the *PKD1* gene. The capability of our methodology to detect this mutation among one of the 20 samples in the pool could support the method validity in terms of sequencing capacity, variant calling and variant filtering.

Concordance of this LR-PCR based targeted MPS and Sanger sequencing: Preliminary results in 2 cases

Completely concordant results were obtained between this LR-PCR based targeted MPS and the previous Sanger sequencing as preliminary evaluation in a representative sample for each gene, PK125 (for *PKD1* gene) and PK150 (for *PKD2* gene). The *PKD1* gene mutation identified in PK125, c.11348_11385delinsAGTCC (p.Val3783_Trp3795delinsGluAla), is a considerably large indel with an insertion of 5 bp together with a 38-bp deletion, while the *PKD2* gene mutation identified in PK150 is just a single-nucleotide substitution, c.2407C>T (p.Arg803Ter). Additional variants which were identified correctly also included several types of base substitutions and different numbers of base duplications. In addition, the design of LR-PCR amplicons which covered most of the introns in *PKD1* and the entire sequence of *PKD2* enabled the detection of deep intronic *PKD2* variants (c.596-3755T>C and c.710-7847C>T) and far-extended 3’ UTR *PKD1* variants (c.*1069_*1072dupTGAC and c.*1080A>G) as well as *PKD1* variants (c.*1999C>T). This reflects the capability of this present method to detect a variety of variant types in both *PKD1* and *PKD2* genes, including those which may be rare mutations in non-coding regions.

In conclusion, this study provided the optimized protocol for targeted MPS of the *PKD1* and *PKD2* genes using LR-PCR amplicons especially in the library preparation step that gave a high coverage of individual samples with high read depths. The modifications included a longer fragmentation time optimal for high-GC content templates which might be the most effective improvement of the protocol. In addition, frequent QC check and PCR-free library preparation also help enhancing the MPS data quality. Trivial modifications to the template preparation protocol even that recommended to be optimal by the manufacturer can enhance the library yield and quality of MPS sequencing data that can lead to a reduction in sequencing costs because of an increased sample multiplexing. This method is valuable for ADPKD molecular diagnostics which require high quality sequencing output.

ACKNOWLEDGEMENTS

This work was supported by Siriraj Research and Development Fund (Type 7, IO: R016138001) as well as Siriraj Core Research Facility (SiCRF), Faculty of Medicine Siriraj Hospital, Mahidol University. We are thankful for the ADPKD patients as well as the Chalermphrakiat Grant, Faculty of Medicine Siriraj Hospital, Mahidol University (CL, KV, MP, WT).

REFERENCES

Barua M, Cil O, Paterson AD, Wang K, He N, Dicks E, Dicks E, Parfrey P, Pei Y (2009) Family history of renal disease severity predicts the mutated gene in ADPKD. J Am Soc Nephrol 20: 1833–1838.

Boucher C and Sandford R. (2004) Autosomal dominant polycystic kidney disease (ADPKD, MIM 173900, *PKD1* and *PKD2* genes, protein products known as polycystin-1 and polycystin-2). Eur J Hum Genet 2: 347–354.

Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T (2015) Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16: 856.

Harris PC, Rossetti S (2010) Molecular diagnostics for autosomal dominant polycystic kidney disease. Nat Rev Nephrol 6: 197–206.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: Overviews and challenges. Biotechniques 56: 61–64.

Huptas C, Scherer S, Wenning M (2016) Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. BMC Res Notes 9: 269.

Katabathina VS, Kota G, Dasyam AK, Shanbhogue AK, Prasad SR (2010) Adult renal cystic disease: a genetic, biological, and developmental primer. Radiographics 30: 1509–1523.

Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D (2011) Systematic comparison of three methods for fragmentation of Long-Range PCR products for Next Generation Sequencing. PLoS ONE 6: e28240.

Kozarewa I, Turner DJ (2011) Amplification-free library preparation for paired-end Illumina sequencing. Methods Mol Biol 733: 257–266.

Lanktree MB, Chapman AB (2017) New treatment paradigms for ADPKD: moving towards precision medicine. Nat Rev Nephrol 13: 750–768.

Liu G, Tan AY, Michaeel A, Blumenfeld J, Donahue S, Bobb W, Parker T, Levine D, Rennert H (2014) Development and validation of a whole genome amplification long-range PCR sequencing method for ADPKD genotyping of low-level DNA samples. Gene 550: 131–135.

Moorthie S, Mattocks CJ, Wright CF (2011) Review of massively parallel DNA sequencing technologies. HUGO J 5: 1–12.

Park ST, Kim J (2016) Trends in Next-Generation Sequencing and a new era for whole genome sequencing. Int Neurourol J 20: S76–S83.

Paul BM, Consugar MB, Lee MR, Sundsbak JL, Heyer CM, Rossetti S, Kubly VJ, Hopp K, Torres VE, Coto E, *et al.* (2014). Evidence of a third ADPKD locus is not supported by reanalysis of designated PKD3 families. Kidney Int 85: 383–392.

Porath B, Gainullin VG, Cornec-Le Gall E, Dillinger EK, Heyer CM, Hopp K, Edwards ME, Madsen CD, Mauritz SR, Banks CJ, *et al.* (2016) Mutations in *GANAB*, encoding the glucosidase IIα subunit, cause autosomal-dominant polycystic kidney and liver disease. Am J Hum Genet 98: 1193–1207.

Qi XP, Du ZF, Ma JM, Chen XL, Zhang Q, Fei J, Wei XM, Chen D, Ke HP, Liu XZ, e*t al.* (2013) Genetic diagnosis of autosomal dominant polycystic kidney disease by targeted capture and next-generation sequencing: Utility and limitations. Gene 516: 93–100.

Ranjzad F, Aghdami N, Tara A, Mohseni M, Moghadasali R, Basiri A (2018) Identification of three novel frameshift mutations in the *PKD1* gene in Iranian families with autosomal dominant polycystic kidney disease using efficient targeted next-generation sequencing. Kidney Blood Press Res 43: 471–478.

Reed B, McFann K, Kimberling WJ, Pei Y, Gabow PA, Christopher K, Petersen E, Kelleher C, Fain PR, Johnson A, *et al*. (2008) Presence of *de novo* mutations in autosomal dominant polycystic kidney disease patients without family history. Am J Kidney Dis 52: 1042–1050.

Renkema KY, Stokman MF, Giles RH, Knoers NV (2014) Next-generation sequencing for research and diagnostics in kidney disease. Nat Rev Nephrol 10: 433–444.

Rossetti S, Strmecki L, Gamble V, Burton S, Sneddon V, Peral B, Roy S, Bakkaloglu A, Komel R, Winearls CG, *et al.* (2001) Mutation analysis of the entire *PKD1* gene: genetic and diagnostic implications. Am J Hum Genet 68: 46–63.

Rossetti S, Chauveau D, Walker D, Saggar-Malik A, Winearls CG, Torres VE, Harris PC (2002) A complete mutation screen of the *ADPKD* genes by DHPLC. Kidney Int 61: 1588–1599.

Rossetti S, Hopp K, Sikkink RA, Sundsbak JL, Lee YK, Kubly V, Eckloff BW, Ward CJ, Winearls CG, Torres VE, *et al*. (2012) Identification of gene mutations in autosomal dominant polycystic kidney disease through targeted resequencing. J Am Soc Nephrol 23: 915–933.

Spithoven EM, Kramer A, Meijer E, Orskov B, Wanner C, Abad JM, Aresté N, de la Torre RA, Caskey F, Couchoud C, *et al* (2014) Renal replacement therapy for autosomal dominant polycystic kidney disease (ADPKD) in Europe: prevalence and survival-an analysis of data from the ERA-EDTA registry. Nephrol Dial Transplant 29: iv15–iv25.

Stevens AJ, Taylor MG, Pearce FG, Kennedy MA (2017) Allelic dropout during polymerase chain reaction due to G-quadruplex structures and DNA methylation is widespread at imprinted human loci. G3 (Bethesda) 7: 1019–1025.

Tan AY, Michaeel A, Liu G, Elemento O, Blumenfeld J, Donahue S, Parker T, Levine D, Rennert H (2014) Molecular diagnosis of autosomal dominant polycystic kidney disease using next-generation sequencing. J Mol Diagn 16: 216–228.

Tan YC, Michaeel A, Blumenfeld J, Donahue S, Parker T, Levine D, Rennert H (2012) A novel long-range PCR sequencing method for genetic analysis of the entire *PKD1* gene. J Mol Diagn. 14: 305–313.

Thomas R, McConnell R, Whittacker J, Kirkpatrick P, Bradley J, Sandford R (1999) Identification of mutations in the repeated part of the autosomal dominant polycystic kidney disease type 1 gene, *PKD1*, by long-range PCR. Am J Hum Genet 65: 39–49.

Thongnoppakhun W, Rungroj N, Wilairat P, Vareesangthip K, Sirinavin C, Yenchitsomanus PT (2000) A novel splice-acceptor site mutation (IVS13-2A>T) of polycystic kidney disease 1 (*PKD1*) gene resulting in an RNA processing defect with a 74-nucleotide deletion in exon 14 of the mRNA transcript. Hum Mutat 15: 115.

Trujillano D, Bullich G, Ossowski S, Ballarín J, Torra R, Estivill X, Ars E (2014) Diagnosis of autosomal dominant polycystic kidney disease using efficient *PKD1* and *PKD2* targeted next-generation sequencing. Mol Genet Genomic Med 2: 412–421.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014a) Ten years of next-generation sequencing technology. Trends Genet 30: 418–426.

van Dijk EL, Jaszczyszyn Y, Thermes C (2014b) Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res 322: 12–20.